Stevens Inst. of Tech.
NsG-494

FACTOR ANALYSIS AND CHEMICAL PROPERTIES

F. B. Clough

I. Introduction.

The technic called factor analysis arises from consideration of sets of variables which are amenable to measurement, with regard to the way in which many replicate measurements of the set yield different values. The different values reflect the different ways in which the variables are determined by the underlying phenomena, that is by the factors. As the multiple factors vary from one sequence of measurements to the next, the variables change in ways which may be quite complicated, according to the magnitudes of the different contributions of each factor to the variables. (Cf. Appendix A.) The analysis we are expounding of measurements of sets of variables, to discover the significant factors, rests first on the supposition that these variables do have common factors* and second on the hypothesis of a linear dependence on the factors.

This is expressed by the equation, for variable k, in replicate measurement i :

$$[1] \qquad P_{ik} - \overline{P}_k = \sum_j n_{ij}\, q_{jk}$$

$\overline{P}_k$ is the mean value of P_{ik} over all replicate measurements. Here we suppose f common factors ($j=1,\cdots,f$). In any particular situation a given factor has a weight expressed by the measure numbers (later called molecular factors, see section II), n_{ij}. n_{ij} in general changes from one

[*] The possibility of unique factors, only relevant to one property, introduces a complication which does not appear to have been completely resolved. [redacted]

replicate measurement to the next, for the same factor j, and we should note that for any other variable, α instead of k, the same n_{ij} appear when the measurement is made under the same environmental conditions: $P_{i\alpha} - \bar{P}_{\alpha} = \sum_j n_{ij} q_{j\alpha}$. The contribution of each of the f factors (subscript j, running from 1 to f) to a given variable is expressed by the values q_{jk}, always the same in every measurement of the variable k . In short, the q_{jk} are <u>the factors</u> for the variable P_{ik}, their measure in a given situation being given by n_{ij}.

The terminology in the above discussion was deliberately chosen to set the point of view for the discussion of chemical properties in the following sections, and the notation represented by equation[1] will be systematically developed for application to chemical properties.

In passing it should be noted that the factors for a set of variables might be chosen in an infinite number of ways unless other considerations such as physical significance enter, but that the <u>number</u> of <u>independent</u> factors is fixed for the set of variables.

II. Factors in Chemical Properties.

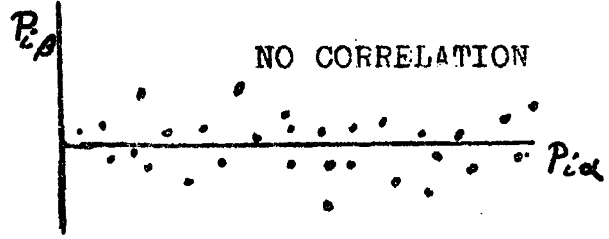
For the consideration of data of chemical interest, it is appropriate to change the terminology to aid in understanding the structure of the problems which we wish to consider from the standpoint of factor analysis.

Our "variables" will be a set of <u>properties</u>, e.g. n_D ; ΔH_v, and so forth, which could be measured on different substances. The "replicate measurements" of these properties, then, are the values of the set obtained for different chemical substances. (We will, following custom, refer to them as measurements on different molecules, even though they may be strictly properties of the aggregate of the molecules, the macroscopic substance.)

This switch in terminology suggests an instructive point of view. When we determine our set of properties on different molecules, we are drawing our samples for our "replicate measurements" out of a large but essentially finite population. If we wish, however, we can consider this a process of sampling an infinite hypothetical population of substances. This population we can think of as being obtained by a continuous variatiom of the parameters of nuclear and electronic charge which determine the properties of matter. The quantization imposed by nature on these properties need not prevent us from taking this statistical point of view. Factor analysis applied to chemical problems, from this point of view, is a kind of analysis of variance.

The objectives of the factor analysis are first to find how many common factors are attributable to a set of properties. Then, if we can find the "measure numbers" n_{ij} (we can refer to these as molecule factors), and the property factors q_{jk} , as in eq.[1], the second objective can be to try to find what combinations of these factors can be given a physical significance. This second objective involves intuitive and speculative consideration of the behavior of matter.

If now there are common factors as we suppose in the set of properties, then of course the properties will be said to be correlated. To begin the analysis of these correlations it is worthwhile to look at the situation for a pair of properties. The presence of several factors in these two properties precludes an analytical expression, $P_{i\beta} = F(P_{i\alpha})$, holding for all molecules i, since (in eq.[1]) the molecule factors n_{ij} will be different for each molecule, giving the factors different weights for each molecule. Instead, a scatter (or correlation) diagram, in which $P_{i\alpha}$ and $P_{i\beta}$ are coordinates, each point showing a pair of values for a particular molecule, may reveal a trend. If there is no trend, hence no correlation, the points will be distributed randomly about one axis, somewhat as shown at the right.



It is apparant that in order to express the degree of correlation a question of scale must be dealt with; the variables must be put in a standardized form.

Also, in order to approach the analysis of multiple correlations, we must take a closer look at the structure of the properties in terms of the factors.

These points will be dealt with in turn.

III. Standardized Variables, and the Correlation Coefficient.

If we regard the measurements of a property for different molecules, as we have suggested, as samplings from an infinite population, it is evident that variables such as $P_{i\alpha}$ or $P_{i\beta}$ should be expressed in terms of the standard deviation of the distribution being sampled. From our sample of this population we estimate the variance σ^2 in the usual way, using the mean for our sample, $\bar{P}_\alpha$. Let us write $x_i = P_{i\alpha} - \bar{P}_\alpha$ and $y_i = P_{i\beta} - \bar{P}_\beta$. In terms of these, the correlation coefficient is defined as (Ref. 1)

$$[2] \qquad \rho_{\alpha\beta} = \frac{\sum_i x_i/\sigma_x \; y_i/\sigma_y}{N} = \frac{\sum_i x_i \, y_i}{N \, \sigma_x \, \sigma_y}$$

The correlation coefficient varies between 0 and 1, depending on the extent of correlation between the variables. It may be observed that the correlation coefficient represents the mean of the regression lines of P_α on P_β and of P_β on P_α. Most important, however, is the recognition that the correlation coefficient also involves the property factors and their measures. In fact, if we make use of eq.[1], we recognize the correlation coefficient to be composed as follows:

$$[3] \qquad \rho_{\alpha\beta} = \frac{1}{N}\sum_i \frac{x_i}{\sigma_\alpha}\frac{y_i}{\sigma_\beta} = \frac{1}{N}\sum_i \frac{P_{i\alpha}P_{i\beta} - \bar{P}_\alpha \bar{P}_\beta}{\sigma_\alpha \, \sigma_\beta} = \frac{1}{N}\sum_i \frac{\left(\sum_j^n n_{ij} q_{j\alpha}\right)}{\sigma_\alpha} \frac{\left(\sum_j^n n_{ij} q_{j\beta}\right)}{\sigma_\beta}$$

In these expressions we see that a kind of averaging over the different molecules, i, has been performed. Consequently the correlation coefficient, whose value can be obtained from the data through expression [2], will be determined primarily by the property factors, $q_{j\alpha}$ and $q_{j\beta}$, provided that the sampling is a good one.

IV. Multiple Correlations of Properties.

A table of the data on a set of properties for a sequence of molecules may be referred to as the data matrix, thus:

$$\begin{array}{l|ccccccccc} & \multicolumn{9}{l}{\text{PROPERTIES (or Variables)}} \\ \hline \text{MOLECULES} & P_{11} & P_{12} & \cdot & \cdot & \cdot & P_{1k} & \cdot & \cdot & \cdot \\ \text{(or Replicate} & P_{21} & P_{22} & \cdot & \cdot & \cdot & \cdot & & \cdot & \\ \text{Determinations} & P_{31} & \cdot & & \cdot & & \cdot & & & \\ \text{of Properties)} & \cdot & & & & & & & & \\ & \cdot & & & & & & & & \\ & P_{i1} & P_{i2} & \cdot & \cdot & \cdot & P_{ik} & & \cdot & \\ & \cdot & & & & & & & & \\ & \cdot & \cdot & & & & \cdot & & \cdot & \end{array}$$

The entire data matrix can be represented as the product of a matrix of property factors, $Q \equiv ((q_{jk}))$, premultiplied by a matrix of the measure numbers characteristic of each factor for each molecule (i.e. the molecule factors, n_{ij}). This is shown in detail in the equation on the following page, where each molecule corresponds to a vector whose components are the molecule factors. This vector times the property factor matrix reproduces one row of the data matrix. In writing this, we wish to fix clearly in mind that it is the elements of the data matrix, $((P_{ik} - \bar{P}_k)) \equiv ((\sum n_{ij}q_{jk}))$, which are the observables, and we wish if possible to discover by analysis of the data both the matrix Q, and the vectors (n_{ij}).

[4]:

Molecule (or replicate measurement)	MOLECULE FACTOR VECTORS xQ (i.e. measure nos. for f factors, ea.molecule)		DATA MATRIX $((P_{ik} - \bar{P}_k))$			
			k=1	k=2		k=p
i = 1	$(n_{11}, n_{12} \ldots n_{1j} \ldots n_{1m}) \times Q$	=	$\sum_j n_{1j} q_{j1}$	$\sum_j n_{1j} q_{j2}$	$\ldots$	$\sum_j n_{1j} q_{jk} \ldots$
i = 2	$(n_{2j}) \quad \times \quad Q$	=	$\sum_j n_{2j} q_{j1}$	.	$\ldots$	
	$\vdots$					
i	$(n_{ij}) \quad \times \quad Q$	=	$\sum_j n_{ij} q_{j1}$	.	. .	$\sum_j n_{ij} q_{jk}$
	$\vdots$					
i = N	$(n_{Nj}) \quad \times \quad Q$	=	$\sum_j n_{Nj} q_{j1}$	.		.

We can restate our problem now as being how to decompose the data matrix into the molecule factor vectors and the property factor matrix, and further how to transform Q, and the molecule factors, so that the factors have a physical significance.

As we have seen, the correlation coefficients between the various properties incorporate a kind of averaging over all of the molecule factors. To study the correlation coefficients systematically, we can form the correlation matrix by premultiplying the data matrix by its transpose. Then the correlation matrix (using data without centering the values on the mean) is:

$$((C_{\alpha\beta})) = \frac{1}{N} ((P_{\alpha i})) \times ((P_{i\beta})) \quad ,$$

and using eq.[1], $P_{ik} = \sum_j n_{ij} q_{jk} + \bar{P}_k$, we obtain eq. [5a] :

(Note: Postmultiplication by the transpose would be used if the data matrix were written with the molecules (replicate measurements) corresponding to columns. This seems to appear in factor analysis literature, but the notation here is consistent with that chosen by Malinowski and Pollara (ref. 2).)

[5a]

$$((C_{\alpha\beta})) = \left(\begin{array}{llll} \frac{(\sum_j n_{ij}q_{j1})^2}{N} + 2\bar{P}_1\sum_i\frac{\sum_j n_{ij}q_{j1}}{N} + \bar{P}_1^2 & \cdot & \cdot & \cdot \\ \cdot & & & \\ \cdot & \text{(general term): } \frac{1}{N}\sum_i(\sum_j n_{ij}q_{j\alpha})(\sum n_{ij}q_{j\beta}) & & \\ & + \frac{1}{N}\bar{P}_\alpha\sum_i(\sum_j n_{ij}q_{j\beta}) & & \\ & + \frac{1}{N}\bar{P}_\beta\sum_i(\sum_j n_{ij}q_{j\alpha}) + \bar{P}_\alpha\bar{P}_\beta & & \\ \cdot & & & \end{array}\right)$$

$$= \left(\begin{array}{llll} \frac{1}{N}\sum_i(\sum_j n_{ij}q_{j1})^2 & \cdot & \cdot & \cdot \\ \cdot & & & \\ \cdot & \frac{1}{N}\sum_i(\sum_j n_{ij}q_{j\alpha})(\sum_j n_{ij}q_{j\beta}) & & \end{array}\right) + 3\begin{pmatrix} \bar{P}_1 \\ \bar{P}_2 \\ \bar{P}_3 \\ \cdot \\ \bar{P}_\alpha \\ \cdot \\ \cdot \end{pmatrix}\begin{pmatrix} \bar{P}_1 & \bar{P}_2 & \cdots\bar{P}_\beta & \cdot & \cdot & \cdot \end{pmatrix}$$

If the properties have been measured from their mean values, and divided by the standard deviation, then we have the reduced correlation matrix $((\rho_{\alpha\beta}))$, instead of $((C_{\alpha\beta}))$. In this the elements are the correlation coefficients as defined in eq.[2].

[5b]

$$((\rho_{\alpha\beta})) = \frac{1}{N}\left(\begin{array}{lllll} 1 & \sum_i(\frac{\sum_j n_{ij}q_{j1}}{\sigma_1})(\frac{\sum_j n_{ij}q_{j2}}{\sigma_2}) & \cdot & \cdot & \cdot \\ & 1 & \cdot\ \cdot\ \cdot & \sum_i\frac{(\sum_j n_{ij}q_{j\alpha})}{\sigma_\alpha}\frac{(\sum'_j n_{ij}q_{j\beta})}{\sigma_\beta} & \\ & \cdot & & \cdot & \\ & \cdot & & & \end{array}\right)$$

Appendix II attempts to clarify the algebra associated with changing to standardized values centered on the mean, and the effect this has on the subsequent treatment of the correlation matrix. The matter is troublesome, at least to the novice.

In viewing eq. 5b (and 6b, p. 11, which will replace it) it is essential to recognize that while the elements of $((\rho_{\alpha\beta}))$ are correlation coefficients directly obtained from the data, it is the molecule and property factors, n_{ij} and q_{jk} which we would like to find. To understand the decomposition of the correlation matrix into the molecule vectors in factor space and the property matrix Q, it is necessary to keep in mind the physical independence of the factors. To this end, a basis for the factor space is conceptually helpful, and this will be introduced in the next section.

V. Factor Space and Independent Factors.

The factor space, in which each molecule is represented as a vector, can be defined in terms of a basis of f orthogonal unit vectors, ε_j.

The reason for introducing an explicit basis is that it helps to carry through the mathematical statement of the physical conception of independent factors. We are supposing for our properties, according to the conceptual picture outlined on p. 3, that varying one factor (in a physical sense) to make a new kind of molecule does not imply varying any other factor.

Since the basis vactors are orthogonal we have $\varepsilon_j \cdot \varepsilon_j = 1$ and $\varepsilon_j \cdot \varepsilon_{j'} = 0$ for $j \neq j'$. The matrix of molecule vectors is now (cf. p. 7)

$$\begin{pmatrix} n_{11}\varepsilon_1 & n_{12}\varepsilon_2 & \cdots & n_{1f}\varepsilon_f \\ n_{21}\varepsilon_1 & n_{22}\varepsilon_2 & \cdots & \\ \vdots & & & \\ & & & n_{Nf}\varepsilon_f \end{pmatrix}$$

Pre-multiplying this by its transpose then leaves all off-diagonal terms zero:

$$\widetilde{((n_{ij}\,\varepsilon_j))}\,((n_{ij}\varepsilon_j)) = \begin{pmatrix} n_{11}^2+n_{21}^2+\dots+n_{N1}^2 & \cdot & \cdot & \cdot & 0 \\ \vdots & \cdot & & & \\ 0 & & \sum_i n_{ij}^2 & & \\ & & & \cdot & \\ 0 & \cdot & \cdot & \cdot & \sum_i n_{if}^2 \end{pmatrix}$$

The elements of the matrix $((q_{jk}))$ are also influenced by the introduction of an orthogonal basis. This matrix, as we have seen on p.7, yields a set of observable properties whenever it is multiplied by the molecule vector $(n_{ij}\varepsilon_j)$ for any molecule i. It contains the independent property factors essential to the property for a particular basis, $\varepsilon_1 \dots \varepsilon_j, \dots \varepsilon_f$.

Now when we form the correlation matrix, its elements in terms of the underlying factors are simplified because of the ortno-normal basis (remembering that $\widetilde{A\cdot B} = \widetilde{B}\cdot\widetilde{A}$, and assuming the $q_{j\alpha}$ already standardized through division by σ_α):

$$((\rho_{\alpha\beta})) = \frac{1}{N}\widetilde{((q_{j\alpha}))}\,\widetilde{((n_{ij}\varepsilon_j))}\,((n_{ij}\,\varepsilon_j))\,((q_{j\beta}))$$

[6a]

$$= \frac{1}{N}\widetilde{((q_{j\alpha}))} \begin{pmatrix} \sum_i n_{i1}^2 & 0 & \cdot & \cdot & \cdot \\ 0 & \sum_i n_{i2}^2 & \cdot & & \\ \cdot & & \cdot & & \\ \cdot & & & \sum_i n_{ij}^2 & \\ \cdot & & & \cdot & \sum_i n_{if}^2 \end{pmatrix} ((q_{j\beta}))$$

Further expansion of this product gives:

[6b]

$$((\rho_{\alpha\beta})) = \frac{1}{N}\begin{pmatrix} q_{11} & q_{21} & q_{31} & \cdots \\ q_{12} & & & \\ q_{13} & & & \\ \vdots & & & \end{pmatrix}\begin{pmatrix} \sum_i n_{i1}^2 q_{11} & \sum_i n_{i1}^2 q_{12} & \cdots \\ \sum_i n_{i2}^2 q_{21} & \sum_i n_{i2}^2 q_{22} & \\ \vdots & & \end{pmatrix}$$

$$= \frac{1}{N}\begin{pmatrix} \sum_i \sum_j n_{ij}^2 q_{j1}^2 & \sum_i \sum_j n_{ij}^2 q_{j1} q_{j2} & \cdots & \cdot \\ \sum_i \sum_j n_{ij}^2 q_{j2} q_{j1} & \sum_i \sum_j n_{ij}^2 q_{j2}^2 & & \\ \vdots & & \ddots & \end{pmatrix}$$

In [6b] the normalization condition is that $\sum_i \sum_j n_{ij}^2 q_{jk}^2 = 1$, and each q_{jk} is understood here to contain the standard deviation, σ_k . Comparison with eq [5b] shows the mathematical simplification arising from the choice of an orthogonal basis for our independent factors.

Vector spaces. Before we turn our attention to the decomposition of the correlation matrix (section VI), we should consider more carefully the role of the factor space for which we have the f vectors, ε_j, as a basis.

Each molecule, as we have seen, is represented by a vector in this space. Also, as eq.[4] shows, each column of the matrix Q is an f-dimensional vector in factor space, and the value of a property — a number — for any given molecule is the scalar product of the molecule vector $(n_{i1} \ldots n_{ij} \ldots n_{if})$ and the property vector

$$\begin{pmatrix} q_{1\alpha} \\ \vdots \\ q_{j\alpha} \\ \vdots \\ q_{f\alpha} \end{pmatrix}$$

How, then, are we to look upon the property vector? Each component of the property vector may be thought of as the mean value of the projections of all of the molecule vectors (all conceivable molecules) onto this particular coordinate, ε_j — this average projection then being multiplied by the particular weight which that factor has in the observed property. The direction of the property vector in factor space is thus determined by the relative significance of the underlying factors in the expression of that particular property.

The scalar product of two property vectors, and hence the angle between them $(\cos \alpha = q_\alpha q_\beta / |q_\alpha||q_\beta|)$ is determined by the correlation coefficient. It is, consequently, a given constant of the physical system. This can be seen by supposing an orthogonal coordinate system of N dimensions, one coordinate for each molecule. If the point representing a property is located in this coordinate system, it would give the property vector as $P_\alpha = P_{1\alpha}\vec{x}_1 + P_{2\alpha}\vec{x}_2 + \cdots + P_{N\alpha}\vec{x}_N$ where $\vec{x}_i$ are orthogonal unit vectors. Then $P_\alpha \cdot P_\beta = \sum_i P_{i\alpha}P_{i\beta}$ = correlation coefficient.

It should be becognized that the <u>rows</u> of the matrix Q are vectors also, this time referred to P coordinates. We can designate these vectors as $^{*}q_j = (q_{j1} \cdots q_{jk} \cdots q_{jP})$ If the properties chosen actually span the factor space, so that they are not independent, then $P \geqslant f$. There are as many vectors $^{*}q_j$ as there are factors. These f independent vectors determine factor space, and as we shall see they are by virtue of the manner of their determination the basis to which the molecule vectors previously were referred.

VI. Decomposition of the Correlation Matrix.

The correlation matrix is formed from the experimental properties of molecules, but our analysis has shown that its elements are determined by the property factors q_{jk} (eq 6b). The terms $\frac{1}{N} \sum_i n_{ij}^2$ which appear in the matrix, summed over all molecules in the sample, must converge on a constant value for each factor j, for good samplings of molecules. Careful statistical terminology distinguishes between the population mean, or variance, and the sample mean or variance, which are estimates of the population parameters. In the same way it is useful to think of a hypothetical infinite population of molecules for which we have $\lim_{i \to \infty} \sum_i \frac{n_{ij}^2}{N}$, and the sample of molecules from which we estimate this population parameter.

We return to expression [6a]. Here the correlation matrix is expressed as a dyadice (ref. 3). This can also be written

$$((\overset{\frown}{q}_{j\alpha}))\begin{pmatrix} \sum_i n_{i1}^2 & 0 & \cdots & \\ 0 & 0 & & \\ \cdot & \cdot & 0 & \cdots \end{pmatrix}((q_{j\beta})) + ((\overset{\frown}{q}_{j\alpha}))\begin{pmatrix} 0 & 0 & \cdots \\ 0 & \sum_i n_{i2}^2 & \cdots \\ \cdot & \cdot & 0 \end{pmatrix}((q_{j\beta})) + \cdots$$

and the second term, for example, then becomes:

$$((\overset{\frown}{q}_{j\alpha}))\begin{pmatrix} 0 & 0 & & 0 & & 0 \\ \sum_i n_{i2}^2 q_{21} & \sum_i n_{i2}^2 q_{22} & \cdots & \sum_i n_{i2}^2 q_{2\beta} & \cdots & \sum_i n_{i2}^2 q_{2P} \\ 0 & \cdot & \cdot & \cdot & & \end{pmatrix}$$

$$= \sum n_{i2}^2 \begin{pmatrix} q_{21}^2 & q_{21}q_{22} & \cdots & q_{21}q_{2\beta} & \cdots & q_{21}q_{2P} \\ q_{22}q_{21} & q_{22}^2 & \cdots & q_{22}q_{2\beta} & & q_{22}q_{2P} \\ \vdots & & & & & \\ q_{2P}q_{21} & q_{2P}q_{22} & \cdots & q_{2P}q_{2\beta} & \cdots & q_{2P}^2 \end{pmatrix}$$

This last matrix is a dyad, the outer product of the vector $\overset{*}{q}_j \equiv (q_{j\beta})$ by itself: $\begin{pmatrix} q_{j1} \\ \vdots \\ q_{j\alpha} \\ \vdots \\ q_{jP} \end{pmatrix}(q_{j1} \cdots q_{j\beta} \cdots q_{jP}) = q_j^{**}q_j$.

In short, from [6a] we have by this argument

$$[7] \quad ((\rho_{\alpha\beta})) = \frac{\Sigma n^2_{i1}}{N} \; q_1^{*}\,{}^{*}q_1 + \frac{\Sigma n^2_{i2}}{N} \cdot q_2^{*}\,{}^{*}q_2 + \ldots \frac{\Sigma n^2_{ij}}{N} \; q_j^{*}\,{}^{*}q_j \; \ldots + \frac{\Sigma n^2_{if}}{N} \; q_f^{*}\,{}^{*}q_f$$

This shows that the correlation matrix is decomposable into a sum of terms, each involving only one factor.

The correlation matrix is geometrically a projection operator (Ref. 3). The vectors ${}^{*}q_j$, we recall, are the rows of the property factor matrix Q. If we further stipulate, as we may, that these vectors are to be an orthonormal set, ${}^{*}q_j \cdot q^{*}_{j'} = 0,\ j \neq j'$, then it is quickly verified that ${}^{*}q_j$ and q^{*}_j are left and right eigenvectors of the correlation matrix, with the eigenvalue $\frac{\Sigma n^2_{ij}}{N}$. Thus:

$${}^{*}q_j((\rho_{\alpha\beta})) = 0 + \ldots + \frac{\Sigma n^2_{ij}}{N} \; {}^{*}q_j \cdot q^{*}_j \, {}^{*}q_j + \ldots = \frac{\Sigma n^2_{ij}}{N} \; {}^{*}q_j \; .$$

The vectors ${}^{*}q_j$ are in fact the basis for the factor space which was introduced in the previous section. Any arbitrary vector in property space can be expanded in terms of an orthonormal basis which includes the vectors ${}^{*}q_j$. When the correlation matrix operates on such a vector, the j<u>th</u> dyad term (in eq. 7) selects the jth component of the vector and multiplies it by $\frac{n^2_{ij}}{N}$; the complete correlation matrix projects the arbitrary vector into factor space, since the eigenvalue is zero for any dimension outside of factor space.

We are now in a position to outline the metuod for decomposing the property matrix into the matrices Q and $((n_{ij}))$, eq. 4, starting with an iteration procedure to isolate the principle factor.

1] Operate on an arbitrary vector ${}^{*}Y$ with the correlation matrix. To see what happens think of this vector expanded in terms of the eigenvectors ${}^{*}q$. Then

$${}^{*}Y((\rho_{ij})) = \lambda_1 y_1 {}^{*}q_1 + \lambda_2 y_2 {}^{*}q_2 + \quad \ldots \quad = {}^{*}Y_I$$

where $\lambda_j = \frac{\Sigma_i n^2_{ij}}{N}$ are the eigenvalues.

Repeated operation by $((\rho_{ij}))$ leads to the dominant factor; i.e. $\frac{{}^*Y_M}{|Y_M|} \longrightarrow {}^*q_{dominant}$, since $\lambda^M_{dom} y_{dom} \gg \lambda^M_j y_j$ for all j other than that corresponding to the dominant eigenvector in the expansion of ${}^*Y$.

2] Form the dyad term $\lambda_{dom}\, q^*_{dom}\, {}^*q_{dom}$ from the eigenvector thus found, and subtract this from the correlation matrix. Repeat the iteration procedure on this reduced matrix to find the second principle factor. Continue in like manner until the new eigenvalue is negligible.

The vectors ${}^*q_j$ thus found form the rows of Q. Using the expression [4], which defines $((n_{ij}))$ and Q, we find that

$$\mathfrak{P}\; P\, \overset{\leftrightarrow}{Q} = ((n_{ij}))\, Q\, \overset{\leftrightarrow}{Q} = ((n_{ij})),$$ by virtue of the orthogonality of the vectors ${}^*q_j$.

Further discussion of computation is deferred to section VIII.

VII. Rotation of the Basis; The Question of Physically Significant Factors.

If it is supposed that the molecular assemblies whose properties are being studied — and which sample the statistical population of all possible variations of the property — can be described in terms of independent physically observable properties, then it should be possible to rotate the basis into these properties. Any transformation of our factor space must preserve the angles between property vectors, since the angles represent the correlations which are the given physical facts of the relations between the variables.

Our search therefore may be for a set of F observable properties whose correlation coefficients are zero, and an orthogonal transformation R which will transform the factors originally arrived at into a pattern in terms of this new basis.

We can express these ideas in terms of the notation which we have developed:

$$[8] \qquad ((n_{ij}))\, R\, R^{-1} Q = ((n_{ij}))\, Q = ((p_{ik}))$$

and since R is an orthogonal transformation, $R^{-1} = \widetilde{R}$. $R^{-1}Q$ is a new property factor matrix whose rows are the new basis vectors.

To develop a procedure for finding R we can begin by writing down the matrix (called a factor <u>structure</u>) which systematically presents the correlations between the molecule factors and the observed properties:

$$S = ((\widetilde{n_{ij}}))((P_{ik})) = ((\widetilde{n_{ij}}))((n_{ij}))Q$$

(Thus an element of S is $S_{jk} = \sum_i n_{ij}P_{ik}$.)

A similar expression can be written for the factors in terms of the rotated basis: $\widetilde{R}S = \widetilde{R}((n_{ij}))((P_{ik}))$. If the values P_{ik} should happen to be just those properties which form the new basis, then it is evident that $\widetilde{R}S$ will have zeros for all elements except those giving the correlation of (P_k) with itself.

We do not need to find a complete rotation at once. We can align one of the original basis vectors with any particular property. Then a second basis vector might be aligned with a property orthogonal to the first (if one is available) and so forth. Each alignment is done by a column of R, which we will represent by R_1^{*}. Then ${}^{*}R_1$ is correspondingly a row of $\widetilde{R}$.

Let X_i represent our "suspected property" for each molecule i, so that we can equate the row vectors: ${}^{*}(X_i) = {}^{*}R_1((\widetilde{n_{ij}}))$.
Hence we can write:

[9a] ${}^{*}R_1 S = {}^{*}R_1((\widetilde{n_{ij}}))((n_{ij}))Q = {}^{*}(X_i)((n_{ij}))Q$

whence ${}^{*}R_1((\widetilde{n_{ij}}))((n_{ij})) = {}^{*}(X_i)((n_{ij})) = \bar{n}$

This gives the relationship for determining ${}^{*}R_1$ from the data:

[9b] ${}^{*}R_1 = \bar{n}\,[((\widetilde{n_{ij}}))((n_{ij}))]^{-1}$

alternatively, $1 = {}^{*}R_1 R_1^{*} = \bar{n}\,[((\widetilde{n_{ij}}))((n_{ij}))]^{-1}R_1^{*}$ and

[9c] $R_1^{*} = [((\widetilde{n_{ij}}))((n_{ij}))]\,\bar{n}^{-1}$